DUAL: Textless Spoken Question Answering with Speech Discrete Unit Adaptive Learning

Anonymous ACL submission

Abstract

Spoken Question Answering (SQA) has gained research attention and made remarkable progress in recent years. However, existing SQA methods rely on Automatic Speech Recognition (ASR) transcriptions, which is time and cost-prohibitive to collect. This work proposes an ASR transcription-free SQA framework named Discrete Unit Adaptive Learning (DUAL), which leverages unlabeled data for pre-training and is fine-tuned by the SQA downstream task. DAUL can directly predict the time interval of the spoken answer from the spoken document. We also release a new SQA benchmark corpus Natural Multi-speaker Spoken Question Answering (NMSQA) for testing SQA in realistic scenarios. The experimental results show that DUAL performs competitively with the cascade approach (ASR + text QA), and DUAL is robust to real-world speech. We will open-source our code and model to inspire more SQA innovations from the community.

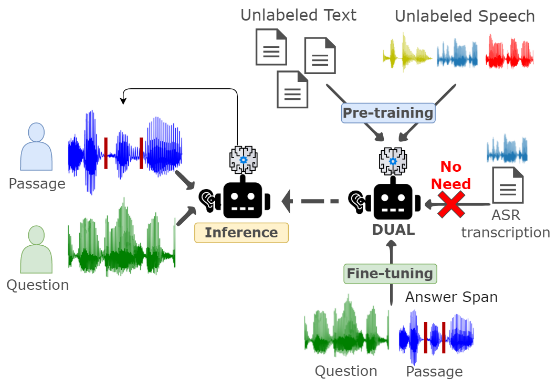

Figure 1: The illustration of the proposed DUAL framework for ASR transcription-free SQA. All the passages, questions, and answers are in spoken form. Our DUAL framework can extract the time interval of the spoken answer from the spoken passage without the help of ASR transcriptions.

1 Introduction

Spoken Question Answering (SQA) aims to find the answer from a spoken document given the question in either text or spoken form. SQA is crucial for speech assistants to answer the question from user spoken queries. The SQA system requires sophisticated comprehension and reasoning ability. It also needs the listening capability to transcribe content from audio. The machine must understand the global and fine-grained information in the spoken context and questions to predict the exact answer span in the long context.

The conventional SQA system consists of an Automatic Speech Recognition (ASR) and a text QA model. However, Lee et al. (2018b) shows that speech recognition errors cause a catastrophic impact on the text QA system. Several works (Lee et al., 2019; You et al., 2021b,c; Su and Fung, 2020) intend to alleviate the negative effect of speech recognition error by knowledge distillation, which is leveraged for adapting the text QA model to be robust against recognition errors. On the other side, Chuang et al. (2020) and Chung et al. (2021) exploit paired speech and transcription to align the semantics for constructing a cross-modal speech and text pre-trained model. The cross-modal model can be fine-tuned end-to-end, mitigating the speech recognition error and improving SQA performance.

Nevertheless, current SQA research suffers from the dependency on ASR transcriptions. It is time-consuming and expensive to collect sufficient transcriptions to train a low error rate and robust ASR. Furthermore, ASR transcriptions are unaffordable in low-resource languages and unavailable in languages and dialects without written form. To make SQA applications more inclusive to all human languages, developing an ASR transcription-free SQA system is critical.

In this work, we propose the first textless (i.e., ASR transcription-free) SQA framework. Inspired by the concept of Textless NLP (Lakhotia et al., 2021; Polyak et al., 2021; Kharitonov et al., 2021;

Kreuk et al., 2021), which encodes speech into the pre-text discrete units, and the surprising findings of pre-trained language model transferability in Kao and Lee (2021), we propose the framework, Discrete Unit Adaptive Learning (DUAL) for textless SQA. DUAL leverages speech pre-trained models to obtain quantized, length-condensed speech representation from continuous audio signals. DUAL further adapts the pre-trained language model for the speech representations and achieves competitive SQA results without any ASR transcriptions.

Furthermore, although there are increasing efforts to build SQA benchmark corpora (Tseng et al., 2016; Lee et al., 2018b; Rajpurkar et al., 2016; You et al., 2020; Lee et al., 2018a; Ravichander et al., 2021), there is a lack of natural SQA datasets to measure SQA performance in environments that capture real-world attributes. To benchmark SQA in a more realistic setting, we release a novel benchmark corpus, Natural Multi-speaker Spoken Question Answering (NMSQA). The corpus has the test set spoken by human readers with text content obtained from in-domain (SQuAD (Rajpurkar et al., 2016)) and out-of-domain (NewsQA (Trischler et al., 2017), QuAC (Choi et al., 2018)) corpora. The training and validation set are synthesized from Amazon Polly TTS service with industrial-grade quality. Different real and synthesized speakers read the pair of (context, question). NMSQA is designed to offer a large-scaled training corpus and human-read testing set for developing and evaluating SQA in real-world scenarios.

Our contributions can be summarized as follows:

- We propose the DUAL framework for SQA, the first work to achieve textless SQA that does not utilize ASR transcriptions.
- We open-source the dataset NMSQA to inspire innovation for SQA in real-world scenarios.
- The experimental results show that DUAL achieves competitive performance and significantly outperforms the cascade approach when the speech recognition error is higher than the 30% word error rate.
- DUAL is more robust to realistic speech than the cascade approach: DUAL retains the performance in the real-speaker testing set, whereas the cascade approach degrades severely.

2 Related Work

Spoken Question Answering: SQA is the crucial use case for voice assistants in our daily life. Currently, there are increasing efforts toward SQA benchmark corpora. TOEFL listening comprehension test (Tseng et al., 2016) is a multiple-choice SQA dataset, but the scale of the data is limited. Spoken SQuAD (Lee et al., 2018b) is the first SQA large-scale dataset. It adopts SQuAD (Rajpurkar et al., 2016) to form a dataset with text questions and spoken documents. Spoken-CoQA (You et al., 2020) is also a large-scale dataset tailored to dialogue SQA. However, they still use the synthetic speech by Google TTS. To push SQA toward real-world scenarios, ODSQA (Lee et al., 2018a) is a large-scale Chinese SQA corpus with real audio recordings. NoiseQA (Ravichander et al., 2021) and SD-QA (Faisal et al., 2021) propose a QA dataset with real spoken question prompts. However, both NoiseQA and SD-QA only contain the spoken queries, and they mainly focus on the text-based QA system. In contrast, our NMSQA dataset includes spoken questions and spoken documents in both naturally synthetic and real speech.

Existing SQA methods intend to improve SQA performance by mitigating or sidestepping the ASR errors. Previous works adopt adversarial domain adaptation (Lee et al., 2019), knowledge distillation (You et al., 2021b,a,c,d), and contextualized word embedding (Su and Fung, 2020) to alleviate the adverse effects of ASR errors. Besides, end-to-end fine-tuning can also ease speech recognition errors. Kuo et al. (2020) tends to fuse acoustic information into the text representation. SpeechBERT (Chuang et al., 2020) and SPLAT (Chung et al., 2021) integrate audio and text information to a joint cross-modal representation for further SQA fine-tuning. Nevertheless, due to the significant disparity between speech and text representation, those cross-modal representations still require ASR transcriptions to align the embedding of speech and text. To the best of our knowledge, existing SQA methods highly rely on ASR transcriptions, and our work is the first step toward transcription-free SQA.

Textless NLP: Recent successes in self-supervised speech representation learning (Hsu et al., 2021; Baevski et al., 2020; Chen et al., 2021; Baevski et al., 2019; Schneider et al., 2019; Riviere et al., 2020; Ling and Liu, 2020; Liu et al., 2021, 2020b,a; Chung et al., 2020, 2019; Ravanelli et al., 2020) enable discovering discrete units from raw waveform without text supervision. The concept of "Textless NLP" is to utilize such discrete units to sidestep the ASR, which needs a large amount of annotated speech and transcription and is only applicable to the written form languages. "Textless NLP" can make speech technologies inclusive to all human languages. Polyak et al. (2021) leverages the discrete units as the content-disentangled component for speech re-synthesis. Lakhotia et al. (2021); Kharitonov et al. (2021) pre-train the speech generative language model based on the discrete units. The speech discrete units can also help the direct speech to speech translation (Lee et al., 2021a,b) and speech emotion conversion (Kreuk et al., 2021). However, previous works of "Textless NLP" focus on speech generation tasks, and our work is centered on the speech semantic task.

Cross-Disciplinary Transfer of Pre-training: Cross-disciplinary transfer refers to transferring knowledge from non-linguistic pre-trained language models (LMs) to natural language or vice versa. Papadimitriou and Jurafsky (2020) show that a non-linguistic data (MIDI music or Java code) pre-trained LSTM-based LM can adapt to natural language LM by only fine-tuning the word embedding. Chiang and Lee (2021) also reveals that even if the language model is not pre-trained on natural languages, the pre-trained models still have the transferability for natural language downstream tasks since the language model learns to model the token dependencies in the sequences. Recently, Kao and Lee (2021) discovered that the text pre-trained models could transfer the learned knowledge to the different downstream tasks of non-text disciplines, such as amino acid, DNA, and music. Specifically, as long as the input sequence is discrete, fine-tuning non-text sequence classification on text pre-trained model yields comparable performance as the non-text data pre-trained model. Since the LMs are pre-trained on a sequential task, the network weights are initialized more sensibly to capture long-range dependencies compared to random initialization schemes. Unlike the previous work, our work is the first to adopt "cross-disciplinary transferability of pre-training" to speech modality.

3 Method

3.1 Problem Formulation

The form of SQA dataset D is $(\mathbf{q}, \mathbf{p}, a)$, corresponding to the passage, question, and answer. $(\mathbf{q}, \mathbf{p}, a)$ is represented in spoken form in this work. Specifically, our goal is to extract the starting and ending time (t_s, t_e), denoted as answer span a, from the spoken passage waveform $\mathbf{p}$ given the spoken question waveform $\mathbf{q}$. Because the output answer is the time interval, the extracted spoken answer is human-recognizable. It does not suffer from speech recognition error or out-of-vocabulary (OOV) as in the case of text answers.

3.2 DUAL framework

The DUAL framework consists of the Speech Content Encoder (SCE) and Pre-trained Language Model (PLM). We introduce the details of the components in the following sections. we illustrate the overview of the DUAL framework in Figure 2.

3.2.1 Speech Content Encoder

The SCE transforms the question-answer audio waveform $(\mathbf{q}, \mathbf{p})$ to sequence of discrete units $(\mathbf{z_q}, \mathbf{z_p})$. The pipeline of SCE is shown in the left part of Figure 2.

Self-supervised Speech Representation: The self-supervised speech pre-trained model can extract informative feature representation. We adopt the state-of-the-art self-supervised speech pre-trained model HuBERT (Hsu et al., 2021) for feature extraction[1]. HuBERT is trained by masked prediction objective similar to BERT (Devlin et al., 2019). The prediction target is the clustering index generated by K-means clustering of signal processing features, e.g., Mel-frequency cepstral coefficients (MFCC) features initially, and then the clustering of learned latent representations in subsequent iterations. We utilize the HuBERT-Large pre-trained model containing 24 transformer encoder layers pre-trained on LibriLight 60k hour dataset. HuBERT encodes the raw waveform into frame-level 1024 dimension features. Each frame is equivalent to 20 ms.

Speech Quantization: The goal of speech quantization is to discretize speech features for feeding discrete units into the pre-trained language model. The K-means clustering is the quantization method, which is trained on the layer-wise representation of

[1]We use the open-source S3PRL (Yang et al., 2021) toolkit to extract HuBERT-Large's representation.

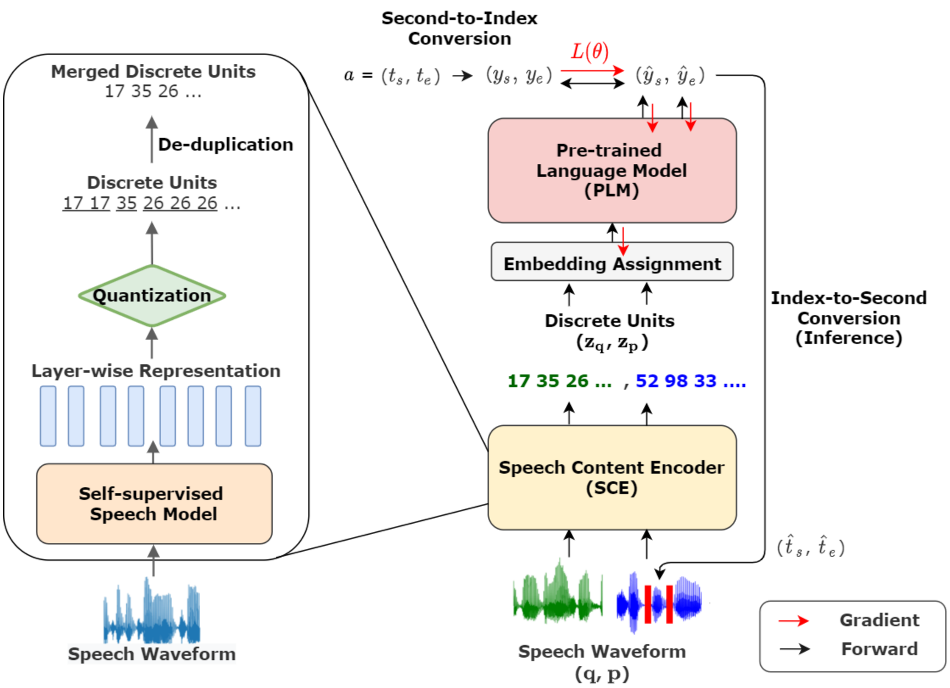

Figure 2: (Left) The pipeline in speech content encoder. (Right) The overview of the DUAL framework.

HuBERT-Large[2]. We use LibriSpeech (Panayotov et al., 2015) 100 hour subset to train the K-means clustering model, and the number of clusters K is 64, 128, and 512. After clustering, the discrete units are represented by the clustering index. We merge the duplicate discrete units to shorten the sequence length and remove the duration information, forming the dense speech discrete sequence of the question and passage $(\mathbf{z_q}, \mathbf{z_p})$. We record the duration of duplication as $\mathbf{c_q}$ and $\mathbf{c_p}$ for $\mathbf{z_q}$ and $\mathbf{z_p}$, so we can recover the frame-level indices to convert the answer span back to time interval at the inference stage.

3.2.2 Pre-trained Language Model

The learning model is a BERT-like transformer encoder model. The input is the discrete units of spoken questions and passages $(\mathbf{z_q}, \mathbf{z_p})$. Because SQA is a very challenging task to train from scratch, we leverage the cross-disciplinary transferability of PLM (Papadimitriou and Jurafsky, 2020; Kao and Lee, 2021; Chiang and Lee, 2021) to help the SQA downstream task. Specifically, we use the weight of text PLM for network initialization and randomly assign the text pre-trained input embeddings for discrete units similar as Kao and Lee (2021). The different random embedding assignments do not significantly affect the final performance (details are in the Ablation Study in Section 5.1). The input of PLM is the concatenated discrete units of question and passage pair $(\mathbf{z_q}, \mathbf{z_p})$, and the target is the start and end position (y_s, y_e) after the de-duplication process.

Because the length of speech discrete units is much longer than text and the duration of the spoken passage itself is long, the standard maximal length of PLM (typically 512) is not enough in our case. As a result, we leverage the sparse transformer PLM for a lengthy document, Longformer (Beltagy et al., 2020), to model the long $(\mathbf{z_q}, \mathbf{z_p})$. Longformer is a BERT-like model for long documents, pre-trained on the unlabeled long text documents and optimized for training efficiency by sparse attention mechanism, such as local and global attention, to support up to 4096 tokens.

3.2.3 Training Objective

The training objective is similar to canonical QA fine-tuning in text QA. A randomly initialized linear layer is added on the top to predict the start and end index. Let θ represents the trainable weights of the model, shown as the gradient flow in Figure 2. $\mathbf{c_p} = [\mathbf{c_{p_1}}, \mathbf{c_{p_2}}, ..., \mathbf{c_{p_n}}]$ is the duration of duplication of every discrete units $\mathbf{z_{p_i}}$ in $\mathbf{z_p} = [\mathbf{z_{p_1}}, \mathbf{z_{p_2}}, ..., \mathbf{z_{p_n}}]$. (t_s, t_e) is the ground truth start and end time in second, and we convert the answer span to index level (y_s, y_e). The overall

[2] We discovered that different layers contain different acoustic and linguistic information. We will discuss this in the ablation study.

Property	train	dev	test-SQuAD	test-OOD
# of Sample	95024	21199	101	166
Hour	297.18	37.61	2.61	8.36
# of Speaker	12	12	60	60
Real Speaker	✗	✗	✔	✔
Content Source	SQuAD-train	SQuAD-dev-1	SQuAD-dev-2	NewsQA-dev, QuAC-dev
Speech Quality	Natural, Clean	Natural, Clean	Disfluent, Noisy	Disfluent, Noisy

Table 1: The properties and splits of NMSQA dataset.

training objective is to minimize the loss $L(\theta)$ as the sum of the negative log probabilities of the true start and end indices on all the examples. $L(\theta)$ can be written as below:

$$-\sum log\, P(y_s|\mathbf{z_q}, \mathbf{z_p}; \theta) + log\, P(y_e|\mathbf{z_q}, \mathbf{z_p}; \theta)$$

At the inference stage, we convert the predicted start and end indices $(\hat{y}_s, \hat{y}_e)$ to the frame level by $\mathbf{c_p}$ and finally transform them to the time level $(\hat{t}_s, \hat{t}_e)$. Since each frame of HuBERT is 20 ms duration, we multiply 0.02 for the second-level time.

$$\hat{t}_s = 0.02 \times \sum_{k=1}^{\hat{y}_s} \mathbf{c_p}_k \quad \hat{t}_e = 0.02 \times \sum_{k=1}^{\hat{y}_e} \mathbf{c_p}_k$$

4 Experiments

4.1 Corpus Description

We propose a new listening comprehension task named Natural Multi-speaker Spoken Question Answering (NMSQA). The details of the NMSQA corpus are listed in Table 1. The train and dev set is the spoken version of the SQuAD v1.1 dataset, one of the largest QA datasets from Wikipedia paragraphs and human-written questions. We randomly split the SQuAD dev set into the disjoint SQuAD-dev-1 and SQuAD-dev-2 for the NMSQA dev set and test set. The Amazon Polly Text-to-Speech service[3] is used for generating natural speech. We randomly assign the 12 TTS speakers and ensure that different speakers speak the spoken document-question pairs. Overall, there are 297.18 / 37.61 hours of audio for the train/dev set.

Moreover, we are releasing two versions of the realistic test set. One is **test-SQuAD**, the human readers are requested to read the SQuAD-dev-2 naturally. Different from test-SQuAD, the **test-OOD** set contains other QA data in NewsQA (Trischler et al., 2017) and QuAC (Choi et al., 2018). Due

[3]https://aws.amazon.com/tw/polly/

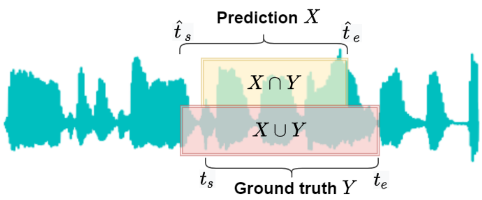

Figure 3: The illustration of evaluation the predicted answer and ground-truth answer in time-level.

to the data distribution shift, the original dev set in NewsQA and QuAC is too tricky for even the text QA model trained on SQuAD. We select the sub-set of development set in NewsQA and QuAC where the text QA model[4] can correctly answer and randomly sample 166 question and answer pairs for human readers. There are 60 human readers, and the gender is balanced (30 males / 30 females). The test-SQuAD and test-OOD have 2.67 and 8.36 hours of audio, respectively. The answer intervals are annotated by force alignment (McAuliffe et al., 2017). The details of human data collection are in Appendix E.

4.2 Evaluation

Since the output target in the NMSQA dataset is the temporal span of the spoken answer, there is no text output to evaluate the Exact Match (EM) or F1 score as in the text QA task. Following the evaluation metric proposed by Lee et al. (2018b) and Chuang et al. (2020), we adopt the Frame-level F1 score (FF1) and Audio Overlapping Score (AOS) to evaluate performance. These two metrics directly measure the SQA performance as a function of the predicted time intervals. The calculation is as follow:

$$P = \frac{X \cap Y}{X} \qquad R = \frac{X \cap Y}{Y}$$

$$FF1 = \frac{2 \times P \times R}{P + R} \times 100\% \qquad AOS = \frac{X \cap Y}{X \cup Y} \times 100\%$$

[4]The text QA model is a typical BERT(*bert-base-uncased*) QA model fine-tuned on SQuAD v1.1 with 88.2 F1 score.

Input	Model	dev		test-SQuAD		test-OOD	
		FF1	AOS	FF1	AOS	FF1	AOS
With ASR transcriptions (Cascade Approach)							
ASR prediction (SB)	Longformer†	56.74	49.72	17.34	15.27	16.92	15.66
ASR prediction (W2v2-st-ft)	Longformer†	65.67	58.34	64.17	57.44	57.67	50.31
Without ASR transcriptions (DUAL)							
HuBERT-64	Longformer	47.76	42.22	39.03	32.97	32.58	28.39
HuBERT-128	Longformer	54.22	48.52	**55.93**	**49.13**	**38.63**	**34.61**
HuBERT-512	Longformer	**55.02**	**49.59**	17.28	12.46	10.35	7.40

Table 2: The performance of DUAL and cascade approach on the NMSQA dev and test set. The Longformer† means the Longformer model has been fine-tuned on clean text SQuAD-v1.1; otherwise, the Longformer is pre-trained by unlabeled text data.

ASR	LibriSpeech test-clean	NMSQA dev	NMSQA test
SB	3.08	15.75	61.70
W2v2-st-ft	1.90	10.48	11.28

Table 3: Word Error Rate (WER) of the two off-the-shelf ASR models on different speech datasets. “NMSQA test” set includes “test-SQuAD” and “test-OOD”.

X is the audio time interval of the predicted answer, and Y is the audio time interval of the ground-truth answer. See Figure 3 for illustration. The higher FF1 and AOS score mean more significant overlapping between the ground truth time interval and the predicted time span.

4.3 Cascade Approach

The SQA cascade approach comprises an ASR model and a QA model trained on clean text. The ASR model is used for Speech-to-Text conversion, and the text QA model will predict the text answer span based on the ASR predictions. The text QA model is a Longformer-based model fine-tuned on SQuAD v1.1, denoted as Longformer† in our experiments. We use the online available model checkpoint[5] for text QA inference. The Longformer† obtains 91.54 F1 score and 85.14 EM score on the text SQuAD v1.1 dataset. Because the final answer target is the time interval of the spoken answer, we adopt the force alignment (McAuliffe et al., 2017) to retrieve the time interval in seconds.

We use two open-source pre-trained ASR models for the cascade approach. One is from Speechbrain (Ravanelli et al., 2021)[6], referred to as **SB**. The other is the open-source Wave2vec 2.0-large with self-training fine-tuning (Baevski et al., 2020)[7], called **W2v2-st-ft** for simplicity. The Word Error Rate (WER) of them on different speech datasets are listed in Table 3, and the details of the ASR models are in Appendix C. Both SB and W2v2-st-ft utilize LibriSpeech (Panayotov et al., 2015) 960 hour dataset as supervised ASR data; however, W2v2-st-ft is much more robust than SB on the NMSQA test set since it leverages the large amount of 60k hour unlabeled data and self-training procedure.

4.4 Results

In the following, “HuBERT-K” means the input of DUAL is the clustering results from HuBERT-Large $22th$ layer representation with K clusters.

4.4.1 Natural and Clean Speech

The natural and clean speech refers to the dev set of NMSQA. The experimental results are shown in Table 2 "dev" column. On the top of the Table 2 are the cascade approaches with paired transcriptions. For the results on the dev set, the performance of ASR prediction with W2v2-st-ft (FF1 65.67) is much better than SB (FF1 56.74) due to its lower speech recognition error. At the bottom part of Table 2, bypassing the ASR transcription stage, DUAL achieves 55.02, 54.22, 47.76 FF1 on the dev set for HuBERT with different discrete codebook sizes 512, 128, 64, respectively. As the size of

[5] https://huggingface.co/valhalla/longformer-base-4096-finetuned-squadv1

[6] https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech

[7] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

the input dictionary discovered from HuBERT representations grows, the performance improves. The performance degradation occurs especially when the unit size is 64, suggesting that small codebook sizes lose important fine-grained content information. On the contrary, using the larger codebook size, e.g., 128 and 512 clusters, can preserve more acoustic information and gain better performance on the dev set. Although DUAL's performance is slightly worse than the ASR cascade model, it is surprising that DUAL's performance is close to the cascade approach (SB). The non-trivial performance of DUAL demonstrates that it learns sophisticated speech semantics directly from speech data without the additional speech-to-text conversion or the supervision of ASR transcription.

4.4.2 Realistic and Noisy Speech

The experimental results are shown in Table 2 "test-SQuAD" and "test-OOD" columns.

In-domain content: The content source of test-SQuAD is in-domain since it is based on SQuAD. We observe that the cascade approach (SB) drops the performance sharply due to the very high WER (61.70) on the real speech, whereas the W2v2-st-ft ASR model is more robust and remains similar performance as in the clean dev set. The evident performance difference for the two cascade approaches reveals the issues of ASR robustness. In reality, training a robust ASR system like W2v2-st-ft with self-training on 60k hours requires huge computational resources not available for many application and research institutions. The undesirable ASR error propagation truly exists in real-world scenarios.

On the other side, when using the appropriate codebook size ($K = 128$), DUAL can retain the performance in test-SQuAD, showing remarkable robustness of realistic speech. The performance on test-SQuAD is even slightly higher than the dev set for HuBERT-128. The surprising robustness may come from the speech quantization and the deduplication procedure, which contains the essential acoustic content information while removing the noise and reducing the impact on disfluent speech (i.e., lots of pauses in speech).

We also observe that the different number of clusters in DUAL causes considerably dissimilar performance. HuBERT-128 obtains 55.93 FF1 score, while HuBERT-64 gets 39.03 FF1 score and HuBERT-512 only attains 17.28 FF1 score. The experimental results indicate that even though the large clustering number stores more acoustic information and gains better performance in a clean dev set, it also amplifies undesirable artifacts in out-of-domain speech and leads to catastrophic domain mismatch. The real-world testing concludes that selecting the adequate clustering number is crucial for robust DUAL performance.

Layer	FF1	AOS
5	35.14	30.49
10	44.89	39.52
15	46.90	41.78
21	51.94	46.59
22	**54.22**	**48.52**
23	53.07	47.63

Table 4: Experiments of clustering on different hidden representations of HuBERT-Large. The number of clusters is 128 for all the experiments. The performance is evaluated on the NMSQA dev set.

Out-of-domain Content: The test-OOD set comes from out-of-domain content sources that differ from SQuAD. Compared to "test-SQuAD", all the performance in Table 2 "test-OOD" drop. The results show that out-of-domain examples are more sensitive to speech recognition errors. The cascade approaches and DUAL both suffer from performance degradation in the test-OOD set.

5 Analysis and Discussion

5.1 Ablation Study

Different Layer for Speech Quantization: Table 4 shows the results of clustering on different layers' hidden representations. We choose the 5, 10, 15, 21, 22, 23 layers for experiments. The best performance is at the $22th$ layer, which achieves 54.22 FF1 score and 48.52 AOS score. The top layers (21, 22, 23) have better performance than the bottom layers (5, 10, 15).

In self-supervised speech representation, different layers encode different acoustic and linguistic information. Chen et al. (2021) shows that HuBERT-Large's top layers contribute most for content and semantic-related tasks (such as Phoneme Recognition and Intent Classification) in the weighted-sum fine-tuning scheme (Yang et al., 2021). Their analysis results align with the experimental results in Table 4, showing that the layers with more content information are more suitable for speech quantization and beneficial to the final SQA performance. We also conduct further layer-wise analysis in Appendix A.

Embedding Assignment	FF1	AOS
Most frequent	54.22	48.52
Least frequent	46.88	41.68
Random	51.66	46.23
Re-emb	8.87	7.23
Scratch (baseline)	6.12	4.91

Table 5: Ablation study of embedding assignment. All experiments use the HuBERT-128 setting. Performance is measured on the NMSQA dev set.

Input Embedding Assignment: Table 5 shows the ablation study of different embedding assignments. "Most frequent" and "Least frequent" mean that we randomly assign the n discrete units to the pre-trained embedding of the top-n and the least-n frequent vocabularies. The vocabulary frequency is determined by the Byte-Pair Encoding (BPE) on unlabeled text data. "Random" refers to randomly selecting pre-trained input embedding regardless of the frequency. "Re-emb" denotes to randomly re-initialize input embedding by the normal distribution. "Scratch" means the Longformer model is not pre-trained on the unlabeled text data.

The experimental results indicate that randomly assigning the pre-trained input embeddings for discrete units does not result in very different performance. The result of the "Random" is comparable to the "Most frequent" initialization, and "Least frequent" causes slightly worse performance than "Most frequent." The performance degradation may come from the poor quality of the least frequent vocabularies' pre-trained embeddings.

5.2 DUAL vs. Cascade Approach

We compare the performance of the cascade approach (SB) and DUAL (HuBERT-128) for different levels of WER. Specifically, we split the NMSQA dev set into subsets by ASR (SB) WER from 0% to 70%. In Figure 4, we observe that the FF1 score drops significantly as the WER rises. This is the typical phenomenon of speech recognition error propagation. On the other hand, DUAL attains a similar FF1 score for different levels of WER sub-groups. Because DUAL does not depend on ASR transcriptions, there is no correlation between WER and DUAL's FF1 score. When the WER is below 30%, the cascade approach outperforms DUAL; but when WER exceeds 30%, DUAL's FF1 score is much higher than the cascade approach. Since the content of SQuAD is based on Wikipedia, it usually includes proper nouns (e.g., abbreviation

Figure 4: Frame-level F1 (FF1) score for DUAL and cascade approach (SB), evaluated on the small groups of full NMSQA dev set at different levels of ASR (SB) WER.

and institution). The Out-Of-Vocabulary (OOV) easily leads to speech recognition error and consequently low SQA performance, whereas DUAL can still retain the performance.

6 Conclusion and Future Work

In this work, we propose the first textless (i.e., ASR transcription-free) SQA framework. The proposed DUAL framework only utilizes unlabeled speech and text data for pre-training and fine-tuning by the spoken questions, passages, and answer time intervals. The DUAL framework directly predicts the answer time span without text supervision or acoustic word boundary. Furthermore, we collected a new natural, multi-speaker SQA benchmark corpus named NMSQA. The NMSQA contains real speakers for the test set and large-scale data for the training and development set. The experimental results show that DUAL performs competitively with the cascade approach on NMSQA. DUAL is also robust to real-world noise in the NMSQA test set when selecting the appropriate codebook size.

We plan to investigate the discrete units pre-training on PLM to capture the better semantic representation of speech for future work. We also want to unfreeze the fixed speech content encoder to fine-tune on SQA jointly.

This work shows proof of concept to model the challenging SQA task by audio-level annotations only. Our DUAL framework is applicable to all spoken languages for building SQA without the supervision of text transcriptions. Furthermore, we hope the NMSQA dataset can help the SQA community develop robust SQA systems.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Cheng-Han Chiang and Hung-yi Lee. 2021. On the transferability of pre-trained language models: A study from artificial datasets. *arXiv preprint arXiv:2109.03537*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin-Shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *INTERSPEECH*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R Glass. 2019. An unsupervised autoregressive model for speech representation learning. In *INTERSPEECH*.

Yu-An Chung, Hao Tang, and James R. Glass. 2020. Vector-quantized autoregressive predictive coding. In *INTERSPEECH*.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2021. SPLAT: Speech-language joint pre-training for spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1897–1907, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Fahim Faisal, Sharlina Keshava, Antonios Anastasopoulos, et al. 2021. Sd-qa: Spoken dialectal question answering for the real world. *arXiv preprint arXiv:2109.12072*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wei-Tsung Kao and Hung-yi Lee. 2021. Is BERT a cross-disciplinary knowledge learner? a surprising finding of pre-trained models' transferability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2195–2208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2021. Textless speech emotion conversion using decomposed and discrete representations. *arXiv preprint arXiv:2111.07402*.

Chia-Chih Kuo, Shang-Bao Luo, and Kuan-Yu Chen. 2020. An audio-enriched bert-based framework for spoken multiple-choice question answering. In *INTERSPEECH*.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021a. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2021b. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee. 2019. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7300–7304. IEEE.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung yi Lee. 2018a. Odsqa: Open-domain spoken question answering dataset. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018b. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *INTERSPEECH*.

Shaoshi Ling and Yuzong Liu. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

Alexander H Liu, Yu-An Chung, and James Glass. 2020a. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020b. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge Set Evaluation for User-Centric Question Answering. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Dan Su and Pascale Fung. 2020. Improving spoken question answering using contextualized word representation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8004–8008. IEEE.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Bo-Hsiang Tseng, Sheng syun Shen, Hung yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *INTERSPEECH*.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020. Towards data distillation for end-to-end spoken conversational question answering. *ArXiv*, abs/2010.08923.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021a. Contextualized attention-based knowledge transfer for spoken conversational question answering. *ArXiv*, abs/2010.11066.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021b. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7793–7797. IEEE.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021c. Mrd-net: Multi-modal residual knowledge distillation for spoken question answering. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3985–3991. International Joint Conferences on Artificial Intelligence Organization.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021d. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 28–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Probing: Content Information for Different Layers

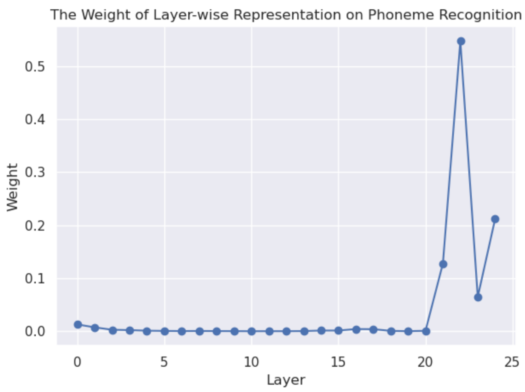

Figure 5: The weight of the final weighted-sum representation is fine-tuned by LibriSpeech 100 hour phoneme recognition downstream task. The most significant weight is at the $22th$ layer.

To investigate the performance gap between the former and the latter layers, we follow the concept of using weighted-sum representation as the final representation to train downstream phoneme recognition (PR) task as in (Yang et al., 2021). By training on LibriSpeech (Panayotov et al., 2015) 100 hour dataset, the frozen Hubert-Large model with trainable weights and upstream linear model achieved 3.53 phoneme error rate (PER) on LibriSpeech test-clean. PR is a content recognition task that transcribes an utterance into the smallest content units (phoneme). The weights of different layers indicate how much content information is stored in that layer. The result is shown in Figure 5. The top layers have significantly larger weight, especially at the $22th$ layer. The results demonstrate that the top layer-wise representation in the HuBERT-Large model encodes more content information than other layers.

B Training Details

For DUAL, we use the official Longformer checkpoint on Longformer-base model[8], which starts from the original RoBERTa checkpoint and is pre-trained for masked language modeling (MLM) on long documents. We search the learning rate in [3e-5, 5e-5, 7e-5, 1e-4] and report the best performance. We set the learning rate warmup step as 500, growing up linearly to the peak value and then linearly decaying to 0. All the DUAL experiments use 4 Tesla V100s with an overall 128 batch size for up to 5000 training steps. The training takes about one day. If the length of discrete units $(\mathbf{z_q}, \mathbf{z_p})$ input exceeds 4096, we truncate the passage $\mathbf{z_p}$.

[8]https://huggingface.co/allenai/longformer-base-4096

C Details of ASR Models

The Speechbrain (**SB**) ASR model consists of CRDNN with CTC/Attention and RNNLM trained on LibriSpeech 960 hour dataset. This ASR model achieves 3.08 WER on LibriSpeech test-clean and obtains 15.75 WER on the development set of the NMSQA dataset but only 61.70 WER on the testing set. The high testing WER points out the ASR robustness issues of the real-world applications.

On the other hand, **W2v2-st-ft** ASR model is the Wav2vec 2.0-Large model. First pre-trained on Libri-light and LibriSpeech, then self-training and fine-tuning on Librispeech 960 hour. W2v2-st-ft achieves 1.90 WER on Librispeech test-clean set. The WER on the NMSQA development and testing set are 10.48 and 11.28, respectively.

D Can we learn sophisticated semantic information solely from speech data?

We try to fine-tune SQA as a downstream task for the state-of-the-art self-supervised pre-trained speech representation model such as HuBERT (Hsu et al., 2021). However, we find out that SQA speech input is too long for self-supervised speech models, which can only receive about 15 seconds of speech; however, the duration of spoken paragraphs is usually longer than 1 minute. The lack of a long-range and efficient self-supervised speech pre-trained model causes the difficulty to model high-level semantic information by speech data itself.

E Details of Human Data Collection

The test set of NMSQA is an audio set collected from human readers reading SQuAD, NewsQA, and QuAC Corpora. The corpora are split into sentences, and human readers are provided content in the form of text sentences and are requested to read and record the audio of the reading. The audio length is around 11 hours, with around 3600 sentences in total that are later composed back to documents. Each sentence is on average 5s or 10 words. The human readers are gender-balanced (30 male, 30 female). For quality control, we had an initial quality control batch of 1.2 hours of audio (425 sentences) by 16 speakers (8 male, 8 female) and evaluated the quality of the initial batch before proceeding the data collection. The recording condition guideline is derived from LibriVox[9] with some adjustments to suit our scenario. A quiet environment is required for recording, and external USB microphones plugged into the computers are preferred to built-in microphones.

For the audio recording, we use the `wav` files (two-channel audio sampled at 44,100 Hz) as the recording format. The readers are advised to a) read the text before recording it, b) allow pauses between sentences and paragraphs, c) enunciate at a relaxed steady pace, d) speak up and try for a steady volume level, e) place the microphone at an appropriate location, f) take breaks in between, to avoid mental and vocal fatigue. The human reader sourcing and data collection are handled by *ANONYMOUS*, a third-party vendor that has established history in data collection for AI and machine learning research. The data collection and storage fully comply with stringent security, privacy, and ethics requirements.

[9] https://librivox.org/